Co-expression networks between protein encoding mitochondrial genes and all the remaining genes in human tissues

João Almeida
i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto.
FEUP – Faculdade de Engenharia da Universidade do Porto.
Porto, Portugal
contact@joaoalmeida.me

Joana Ferreira
i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto
Porto, Portugal
joanaf@ipatimup.pt

Rui Camacho
Departamento de Informática FEUP – Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
rcamacho@fe.up.pt

Luisa Pereira
i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto
Porto, Portugal
lpereira@ipatimup.pt

***Abstract*— Recent advances in sequencing allow the study of all identified human genes (≈ 22,000 protein encoding genes), which have differential expression between tissues. However, current knowledge on gene interactions lags behind, especially when one of the elements encodes a mitochondrial protein (≈ 1500). Mitochondrial proteins are encoded either by mitochondrial DNA (mtDNA; 13 proteins) or by nuclear DNA (nDNA; the remaining), which implies a coordinated communication between the two genomes. Since mitochondria coordinate several life-critical cellular activities, namely energy production and cell death, deregulation of this communication is implicated in many complex diseases such as neurodegenerative diseases, cancer and diabetes. Thus, this work aimed to identify high co-expression groups between mitochondrial genes-all genes, and associated protein networks in several human tissues (Genotype-Tissue Expression database). We developed a pipeline and a web tree viewer that is available at GitHub (https://github.com/Pereira-lab/CoExpression). Biologically, we confirmed the existence of highly correlated pairs of mitochondrial-all protein encoding genes, which act in pathways of functional importance such as energy production and metabolite synthesis, especially in brain tissues. The strongest correlation between mtDNA genes are with genes encoded by this genome, showing that correlation among genes encoded by the same genome is more efficient.**

Keywords*— *co-expression protein networks, nuclear genome, mitochondrial genome, human tissues, BioTree Viewer

I. INTRODUCTION

The public access to big genomic and transcriptomic data allows detailed omics studies and creates new bioinformatic challenges [1]. The study of these data shed light on the genomic influence in the human phenotype and on understanding complex diseases [2].

The human genome is organized in nuclear genome (nDNA) localized in the nucleus, and in mitochondrial genome (mtDNA) placed inside the cytoplasmic organelles called mitochondria. Mitochondria are responsible for the production of the cell energy and, so, they play an essential role in the cellular life. Their malfunction has been associated with complex diseases such as cancer, diabetes, neurodegenerative disorders, etc [3]. mtDNA codes 13 mitochondrial proteins, but 99% of the known mitochondrial proteins are coded by nuclear genes (around 1500) [4]. The two genomes must be coordinated through mechanisms which are still unknown [4]. Thus, it is important to investigate how nDNA and mtDNA genes correlate in the various tissues of the organism.

The public availability of databases as GTEx [5] allows researchers to address this issue. This database contains data on gene expression of all human genes (protein and non-protein) in several tissues extracted from healthy humans who died mainly from accidents.

II. AIMS

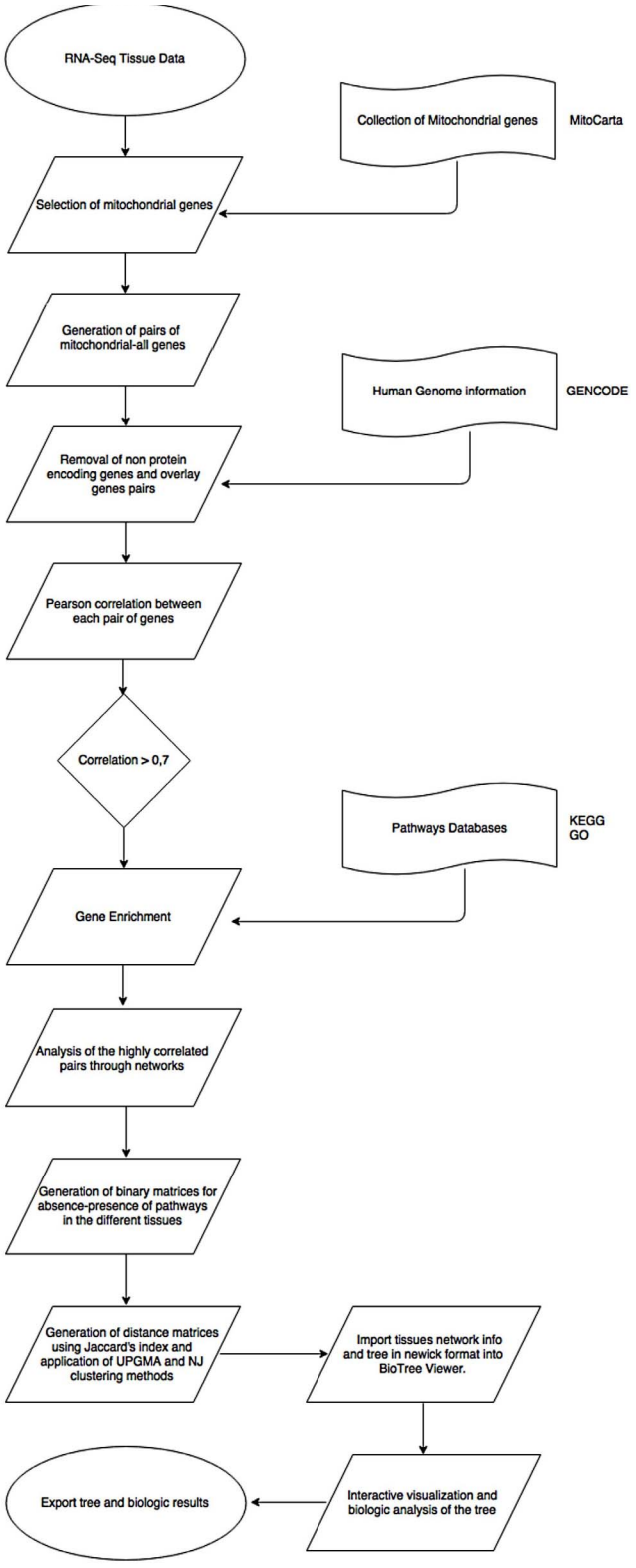
This work aims to identify high co-expression groups between genes encoding mitochondrial proteins and genes encoding all proteins in human healthy tissues, in order to understand the nDNA and mtDNA coordination in phenotypic expression of the individual. A generic pipeline was built (Figure 1) and tested in this work. This pipeline can be used in the future, in similar analysis in other organisms or in tumor presence situations.

III. WORK DESCRIPTION

Gene expression data for tissues were collected from the Genotype-Tissue Expression database (https://www.gtexportal.org/home/) in a total of 49 tissues (8527 samples, an average of 174 per tissue) [5]. RNA-Seq information was processed by a parser which organized data in a specific structure.

By using MitoCarta database information (a collection of 1158 protein encoding mitochondrial genes) and gencode project information, the data was filtered to include only protein-encoding and physically nonoverlapping genes (only one of the overlapping genes was maintained). Previously to the Pearson correlation calculation, expression values that were not in the range of $[u_x - 4SD,\ u_x + 4SD]$ or $[u_y - 4SD,\ u_y + 4SD]$ (SD stands for standard deviation) per gene in each tissue

were considered outliers and thus excluded from data to be analyzed. Gene pairs with a correlation higher than 0.7 were saved to be analyzed.



exported to construct binary matrices:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & a_{3n} \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, (a_{ij}) \in \{0,1\}^{m*n} \quad (1)$$

where m is the total number of tissues available in the analysis and n the number of pathways bearing highly correlated genes included in the biologic database.

For each matrix a similarity matrix of pathways bearing highly correlated genes was built using Jaccard's index through the vegdist method from vegan library in R environment. The matrices were then used as source of data for hierarchical clustering techniques, Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [9] using hclust method and Neighbor Joining (NJ) [10] using nj method from ape library. As result, trees were generated both in newick and image formats.

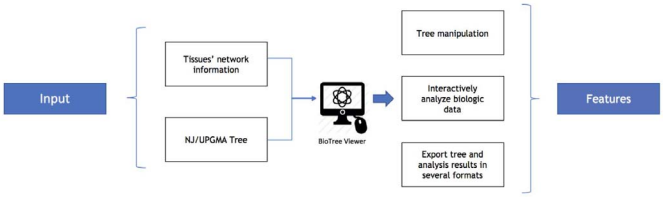


Fig. 2. Diagram of the BioTree Viewer platform.

A web platform was built to preview and analyze interactively the biologic data from the tissue trees (Figure 2). Although this platform is based on phylotree.js software for tree visualization, all the biologic features and the importation and exportation of data were created on the scope of this work, making this platform unique in comparison with what is currently available. The data to feed into the platform can be retrieved through the pipeline previously addressed, from Cytoscape networks. The datasets were mapped and organized by tissue names and by the different pathways included in the biologic databases which contain gene sets representing the different interactions in the networks.

Thus, by using this platform, it is easy to verify if a tissue co-expression network is enriched for a certain pathway and which genes are included in the interactions.

IV. Results

A. BioTree Viewer platform

The BioTree Viewer web platform developed here (Figure 3) requires a Javascript compatible web browser, and it incorporates three components: (1) import of trees in newick format, visualization and manipulation; (2) import of biological information and its visualization in an interactive way in the tree; (3) export of the tree (.png format) and of the biological data (.xls or .csv formats).

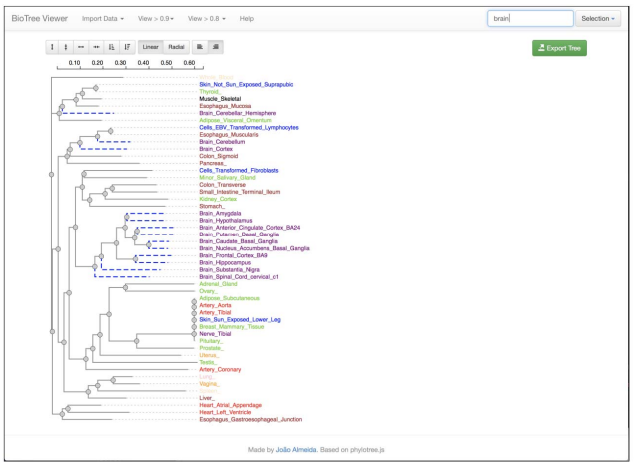


Fig. 3. Visualization of a tree at the BioTree Viewer platform. The colors of tissues represent systems of tissues.

The real novelty of this tool is the possibility of imputing biological data and allowing, in an interactive way, its analysis, visualization and export. By clicking on a node of the tree, the user can check which pathways and genes are shared between all tissues descending from that node (Figure 4).

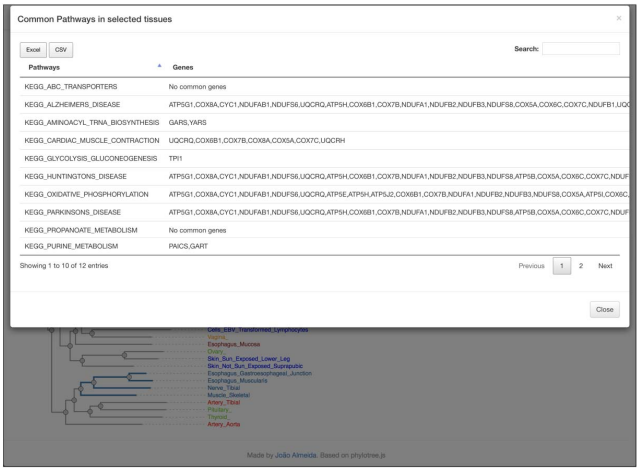


Fig. 4. BioTree Viewer platform information of shared pathways and genes between tissues descending from a node of the tree.

B. Biological results

After filtering the correlation values higher than 0.7 the amount of data was still too big to process and we decided it would be useful to filter even more the data to correlations higher than 0.8 and 0.9.

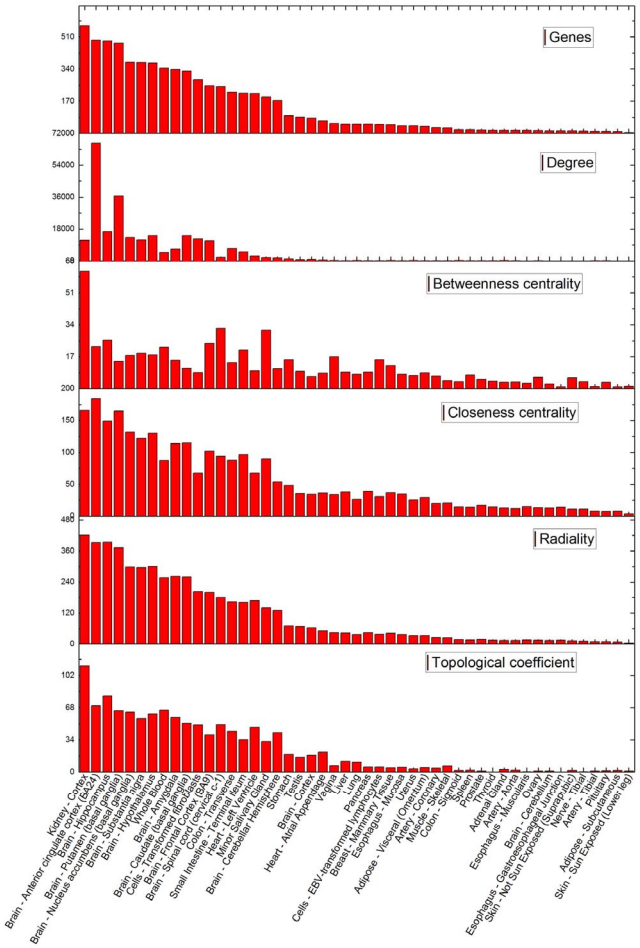


Fig. 5. Complexity parameters of the networks of genes with correlations>0.9 in the 49 human tissues, infered on the Cytoscape [8] tool.

In networks of genes correlated >0.9 the brain tissues (with the exception of brain-cerebellum), together with kidney-cortex, whole blood and cells-transformed fibroblasts present globally more mitochondrial genes (between 16% and 33% of total mitochondrial genes) highly correlated with other genes (left side of graphs in Figure 5, which represents some parameters of the network of correlated genes). In the brain, genes present a high degree of correlations, so networks are extremely dense, while in kidney-cortex the 570 correlated mitochondrial genes (38%) have a lower number of correlations, forming independent clusters and raising the betweenness centrality value.

When focusing on genes correlated >0.8, we verified that between 70%-86% of the mitochondrial genes are highly correlated with other genes in the brain tissues, and in kidney-cortex, whole blood and minor salivary glands, while in most of the other tissues that value varies between 18%-52%.

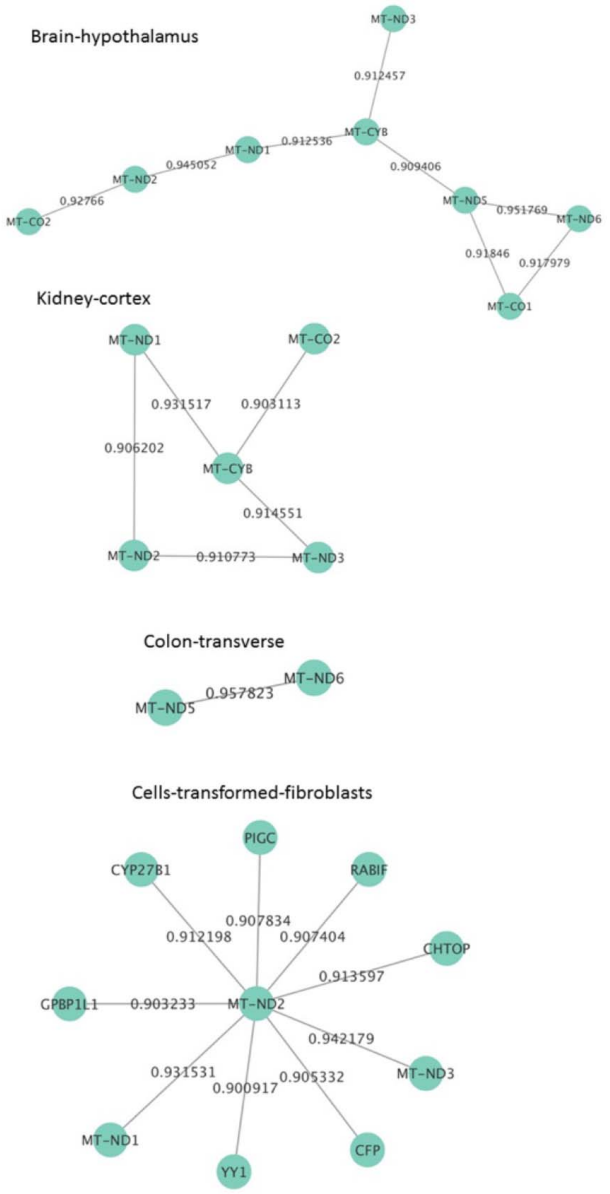


Fig. mtDNA (genes identified by suffix MT). These correlations were only present in brain-hypothalamus, colon-transverse, kidney-cortex and cells-transformed-fibroblasts.

We verified that the mtDNA genes are mostly highly correlated with genes coded by the same genome (Figure 6). Overall, tissues with high energy demanding seem to have a higher coordination of genes involved in energy production, but the two genomes appear to be quite independent.

V. Conclusions

BioTree Viewer application was developed and tested by professionals in the genomics area. The biologic analysis of trees generated by UPGMA and NJ was done by using this platform proving it can have great utility.

The platform and pipeline developed in this work can be adopted for visualization and data study of similar analysis, conducted in other species or in the context of human diseases.

The computational time for the correlation between genes is currently quite long, but can be trivially parallelized. We plan to produce a parallel version in the next release of the tool.

References

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–51, 2001.

[2] L.-L. HSIAO, F. DANGOND, T. YOSHIDA, R. HONG, R. V. JENSEN, J. MISRA, W. DILLON, *et al.*, "A compendium of gene expression in normal human tissues," *Physiol. Genomics*, vol. 7, no. 2, pp. 97–104, Dec. 2001.

[3] S. DiMauro and E. A. Schon, "Mitochondrial DNA mutations in human disease," *Am. J. Med. Genet.*, vol. 106, no. 1, pp. 18–26, 2001.

[4] P. F. Chinnery, "Mitochondrial DNA in Homo Sapiens," in *Human Mitochondrial DNA and the Evolution of Homo sapiens*, H.-J. Bandelt, V. Macaulay, and M. Richards, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 3–15.

[5] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, *et al.*, "The Genotype-Tissue Expression (GTEx) project," *Nat. Genet.*, vol. 45, no. 6, pp. 580–585, May 2013.

[6] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. Database issue, pp. D353–D361, Jan. 2017.

[7] T. G. O. Consortium, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[8] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, *et al.*, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[9] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *Univ. Kansas Sci. Bull.*, vol. 38, no. 2, pp. 1409–1438, 1958.

[10] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, Jul. 1987.